HAL
archives-ouvertes.fr

Automated flow cytometric analysis across large numbers of samples and cell types.

Xiaoyi Chen, Milena Hasan, Valentina Libri, Alejandra Urrutia, Benoît Beitz, Vincent Rouilly, Darragh Duffy, Étienne Patin, Bernard Chalmond, Lars Rogge, et al.

▶ **To cite this version:**

Xiaoyi Chen, Milena Hasan, Valentina Libri, Alejandra Urrutia, Benoît Beitz, et al.. Automated flow cytometric analysis across large numbers of samples and cell types.. Clinical Immunology, Elsevier, 2015, 157 (2), pp.249-60. <10.1016/j.clim.2014.12.009>. <pasteur-01341702>

HAL Id: pasteur-01341702

https://hal-pasteur.archives-ouvertes.fr/pasteur-01341702

Submitted on 4 Jul 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Clinical Immunology (2015) 157, 249–260

available at www.sciencedirect.com

Clinical Immunology

www.elsevier.com/locate/yclim

Automated flow cytometric analysis across large numbers of samples and cell types

Xiaoyi Chen [a,b], Milena Hasan [c], Valentina Libri [c], Alejandra Urrutia [c,d,e], Benoît Beitz [c], Vincent Rouilly [f], Darragh Duffy [c,d,e], Étienne Patin [i], Bernard Chalmond [g,h], Lars Rogge [c,i], Lluis Quintana-Murci [j,k], Matthew L. Albert [c,d,e,l,*], Benno Schwikowski [a,**] for the Milieu Intérieur Consortium

[a] Systems Biology Lab, Institut Pasteur, Paris, France
[b] Laboratory of Analysis Geometry and Modeling, Department of Mathematics, University of Cergy-Pontoise, Ile de France, France
[c] Center for Human Immunology, Institut Pasteur, Paris France
[d] INSERM U818, France
[e] Laboratory of Dendritic Cell Immunobiology, Department of Immunology, Institut Pasteur, Paris France
[f] Center for Bioinformatics, Institut Pasteur, Paris France
[g] University of Cergy-Pontoise, France
[h] CLMA, ENS-Cachan, France
[i] Laboratory of Immunoregulation, Department of Immunology, Institut Pasteur, Paris France
[j] Unit of Human Evolutionary Genetics, Department of Genomes & Genetics, Institut Pasteur, Paris, France
[k] CNRS URA3012, France
[l] INSERM UMS20, France

Received 4 July 2014; accepted with revision 20 December 2014
Available online 7 January 2015

KEYWORDS
Flow cytometry;
Multidimensional analysis;
Population-based cohort;
Automation;
Standardization;
Algorithms;

Abstract Multi-parametric flow cytometry is a key technology for characterization of immune cell phenotypes. However, robust high-dimensional post-analytic strategies for automated data analysis in large numbers of donors are still lacking. Here, we report a computational pipeline, called FlowGM, which minimizes operator input, is insensitive to compensation settings, and can be adapted to different analytic panels. A Gaussian Mixture Model (GMM)-based approach was utilized for initial clustering, with the number of clusters determined using Bayesian Information Criterion. Meta-clustering in a reference donor permitted automated identification of 24 cell types across four panels. Cluster labels were integrated into FCS files, thus permitting

Abbreviations: BIC, Bayesian Information Criterion; CV, coefficient of variation; DC, dendritic cell; EM, Expectation Maximization; FSC, forward scatter; GMM, Gaussian Mixture Model; MFI, mean fluorescent intensity; SSC, side scatter

* Correspondence to: M. L. Albert, Unit of Dendritic Cell Immunobiology, Inserm U818, Institut Pasteur, 25, Rue du Dr. Roux, 75724 Paris Cedex 15, France. Fax: +33 1 45 68 85 48.
** Correspondence to: B. Schwikowski, Systems Biology Lab, Institut Pasteur, 25, rue du Dr. Roux, 75724 Paris Cedex 15, France. Fax: +33 1 40 61 37 01.
E-mail addresses: albertm@pasteur.fr (M.L. Albert), benno@pasteur.fr (B. Schwikowski).

http://dx.doi.org/10.1016/j.clim.2014.12.009
1521-6616/© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

comparisons to manual gating. Cell numbers and coefficient of variation (CV) were similar between FlowGM and conventional gating for lymphocyte populations, but notably FlowGM provided improved discrimination of "hard-to-gate" monocyte and dendritic cell (DC) subsets. FlowGM thus provides rapid high-dimensional analysis of cell phenotypes and is amenable to cohort studies.
© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Flow cytometry is a key technology for the characterization of the cellular component of the immune system. Flow cytometers are able to simultaneously quantify different surface markers of single cells, allowing the identification and quantification of different immune cell subpopulations. In recent years, improvements in measurement speed and experimental automation have enabled comprehensive immunoprofiling of larger cohorts [1].

The gold standard for the analysis of raw flow cytometry data has until now remained "hand gating" (i.e., analysis through computer-assisted procedures for the classification of cells into single cell types using software tools such as FlowJo [2]). Each sample is analyzed by successively separating cell types by successive "gating" in a series of one- or two-dimensional projections. However, the manual operation is laborious and subject to biased visual inspection and gate adjustment. These concerns grow with increased numbers of measured phenotypic markers. Moreover, there is a major limitation in that information critical for accurate gating may not be present in the selected two-dimensional projections.

Here, we report a new method for analyzing multi-parametric flow cytometry, the need for which was motivated by the Milieu Intérieur study. This project aims at defining the genetic and environmental determinants of variable immunologic phenotypes in a healthy population [Thomas et al., co-submission]. Cell phenotyping constitutes one of the major data sets to be integrated into the data warehouse, and as such efforts were made to standardize each step of the sample collection, technical procedures and data analysis. A Companion paper highlights the pre-analytical semi-automated measures put in place for labeling and data generation [Hasan et al., co-submission]. This manuscript details the automated analytic workflow developed for the identification and analysis of 24 cell types across four 8-color cytometry panels.

Our work follows from a large number of computational approaches that have been developed for automated flow cytometry analysis. Recently, the FlowCAP study evaluated a range of approaches [13]. In all cases, however, the datasets used by these investigators were of a smaller scale than the ones in our study, in terms of samples studied (FlowCAP: up to 30 samples; here: 115 samples × 4 panels), and the number of events per experiment (FlowCAP: up to approximately 100,000 events; here: on average 300,000 events per FCS file). Due to these differences, we found that top-ranked FlowCAP approaches were inadequate to address the needs of our data sets. For example, the ADICyt approach [4] required more than 6 h for the analysis of a single sample. The FlowMeans software [5] was faster, but required manual assignment of cell types to each cluster in every single sample. The recent X-Cyt approach [3] was designed explicitly to efficiently address the problem of larger numbers of samples. However, X-Cyt still requires the definition of a "partitioning scheme", a series of mixture models whose sequence and parameters have to be manually configured and calibrated for each cell type of interest in any given analytic panel.

To support the analysis of the Milieu Intérieur cohort dataset, we developed a novel high-dimensional data analysis approach, which we refer to as FlowGM, utilizing fast algorithms that enable the standardized analysis of large numbers of samples. We describe its application to two representative 8-color panels with up to 11 cell populations classified per panel. Its principal feature is that, after the definition of global parameters in a reference sample (i.e., a one-time manual assignment of cell type labels to clusters), it is possible to automatically position and identify cell populations across the entire dataset. This approach will enable analysis of our large healthy donor cohort.

2. Materials and methods

2.1. Dataset

Four 8-color cytometry panels targeting major leukocyte populations across 115 individuals from different age groups and genders were designed to characterize the major immune cell populations (T cells, B cells, NK cells and monocytes), as well as subpopulations of T cells, dendritic cells (DC) and polymorphonuclear leukocytes (PMN). The standardized procedure of collection and treatment of the whole blood sample is described in [Hasan et al., co-submission]. For each of the four panels, technical replicates performed by five parallel blood samples obtained from three donors ("repeatability" studies from [Hasan et al., co-submission]) were generated to examine robustness of the experimental and computational protocols.

2.2. FlowGM cluster model

The input to FlowGM is a set of m sets of n quantitative measurements ("events"), formally, m n-dimensional vectors. Clustering is based on a multivariate Gaussian Mixture Model (GMM) [6], which has the form

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{j=1}^{k} \alpha_j \, \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)$$

A GMM thus corresponds to a set of k clusters, each described by a cluster weight α_j and an n-dimensional

Gaussian (normal) probability distribution, whose parameters $\boldsymbol{\theta}$ are its centroid $\boldsymbol{\mu_j}$, and its extent and orientation, $\sum_j$ in n dimensions. The weight of each cluster corresponds to the proportion of all cells assigned to it. Gaussian mixture models have been used for flow cytometry, but a particularity of FlowGM is that several such clusters can be used to model cells of one type that may not adequately be modeled by a single normal distribution.

2.3. Clustering cells using Expectation Maximization (EM)

Starting from an initial configuration, the degree of fit between the clusters and the data is quantified by a likelihood function. Each stage of an iterative optimization process (Expectation Maximization, EM) improves the likelihood in two steps [7]. In an E (Expectation) step, each event is assigned to (potentially, multiple) clusters whose location is close to the event. In an M (Maximization) step, the cluster parameters are optimized to fit the events assigned to it.

2.4. FlowGM workflow

Step 1 Define pre-processing parameters (manual)
To initialize automatic processing of Phase I, FlowGM requires the input of a few parameters, such as the choice of a reference sample, and the selection of potential pre-filtering and post-filtering parameters.

Step 2 Perform pre-filtering (automatic)
Automated pre-filtering helps eliminate noise (such as doublets) and/or "uninteresting" cells (i.e., Dump populations), which is of importance when the cell types of interest are rare. Two filters have been pre-configured: A doublet filter and a filter that eliminates cells that are negative relative to user-definable markers (based on two-component one- or two-dimensional GMMs). The filter eliminates the 95th percentile of the cluster corresponding to the "uninteresting" cells.

Step 3 Determine the number of clusters (automatic)
The number of clusters (k) used to model the reference sample is determined by minimizing the Bayesian Information Criterion (BIC) [8]. The BIC represents a tradeoff maximizing the degree of fit between the cluster model and the data on the one hand (expressed by the likelihood $\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{\theta})$), and, minimizing, on the other hand, model complexity (based on the number of clusters k):

$$\mathbf{BIC}_{\boldsymbol{k}} = -2\ln(\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{\theta})) + \boldsymbol{k}\ \ln(\boldsymbol{m}).$$

Specifically, we choose k that minimizes the average of $\mathbf{BIC}_k$ under 20 EM runs starting with random initial configurations.

Step 4 Establish the reference clustering (automatic)
Once the number k of clusters has been determined, FlowGM determines 100 random initial configurations of k clusters as starting points, and performs clustering using Expectation Maximization, as described in Section 2.4. The resulting clustering with the highest likelihood is selected as the reference clustering in the second FlowGM phase.

Step 5 Label reference clusters with cell types (manual)
An operator defines the cell types of interest, and assigns one or more corresponding clusters to each such cell type (*labeling*). Thus, each cell type of interest corresponds to a set of clusters (*meta-cluster*).

Step 6 Perform post-filtering (automatic, optional)
This optional step offers the possibility of eliminating additional "uninteresting" events that remain in the clusters determined in Step 5 (analogous to a "dump" gate for conventional approaches and useful in focusing the clustering analysis). Two filters have been pre-configured: A dead cell filter (based on the Viability channel), and a "dump" filter that eliminates selected cells in specified meta-clusters. In both instances, the cells above or below a defined threshold are removed. This threshold is automatically determined as the 95th/99th percentiles of a fitted one-dimensional Gaussian distribution of a reference population along a pre-defined channel. The reference population may either be the meta-cluster itself, or a negative control that has been removed in the pre-filtering (Step 2).

Step 7 Cohort samples: pre-filter and cluster by adjusting labeled reference clustering (automated)
After the reference sample has been processed in Steps 1–5, FlowGM processes all other samples in a fully automated manner. Pre-filtering proceeds as described for the reference donor (Step 2). FlowGM then determines the clustering using EM, as described in Section 2.4, starting with the labeled reference clustering (from Step 5) as the initial configuration. Finally, post-filtering is applied (if selected), as in Step 6.

2.5. Visualization of the resulting clusters in FlowJo

One innovation incorporated into FlowGM included the embedding of labels for each cluster and meta-cluster as additional attributes (numerical identifiers) for each cell in the FCS data file. This allows inspection of the different clusters in FlowJo [2] or other software that can analyze FCS data files.

2.6. Software implementation

FlowGM was implemented using Matlab and Statistics Toolbox Release 2012b [9] and R (version 3.0.1) [10] flowCore package [11]. The visualization graphs were prepared with FlowJo software version 9.7.5.

3. Results

3.1. FlowGM workflow

Motivated by the need for high-quality analysis of a large flow cytometry data set, we developed the novel, and largely automated FlowGM data analysis approach. Its computational high-dimensional clustering approach avoids the limitations inherent to analysis based on two-dimensional projections (Fig. 1A). Experimental data is modeled as a mixture of

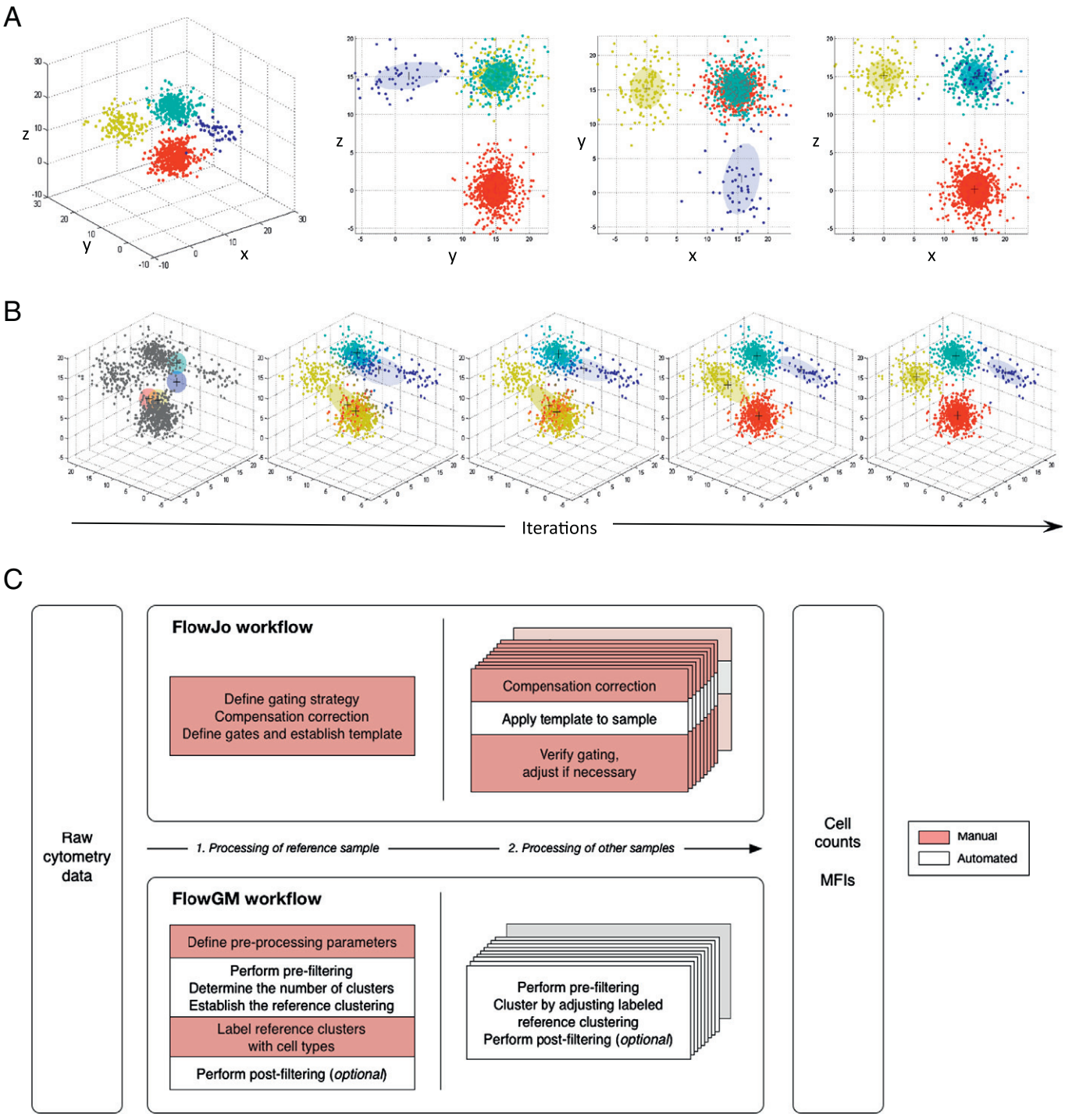


Figure 1 Analytic approach for multidimensional clustering of multi-parameter cytometric data. (A) Four simulated clusters in 3D space that cannot be separated in any 2D projection. (B) Illustration of the expectation–maximization (EM) clustering algorithm using Gaussian mixture model (GMM) clusters, when applied to this data. Points are colored according to their posterior likelihood, the ellipsoid reflects cluster shape, '+' indicates the cluster centroid, transparency of each ellipsoid reflects cluster weight. Five phases are shown: initial random parameter values, updated parameters after the first M-step, after two iterations, after ten iterations, and final solution. (C) FlowJo and FlowGM workflows.

normal distributions (See Materials and methods, Section 2.3) and employs Expectation Maximization (EM) to iteratively adapt model parameters (Fig. 1B and Materials and methods, Section 2.4).

The overall operation of the FlowGM workflow can be understood on the basis of its similarities and differences relative to the current 'gold standard' manual FlowJo workflow (Fig. 1C). For both approaches, two phases can

be distinguished. In the first phase, method parameters are calibrated on selected reference samples. In a second phase, all other samples are processed on the basis of the calibrated parameters. To be suitable for large cohort studies, FlowGM was designed to minimize the manual per-sample effort in the second phase.

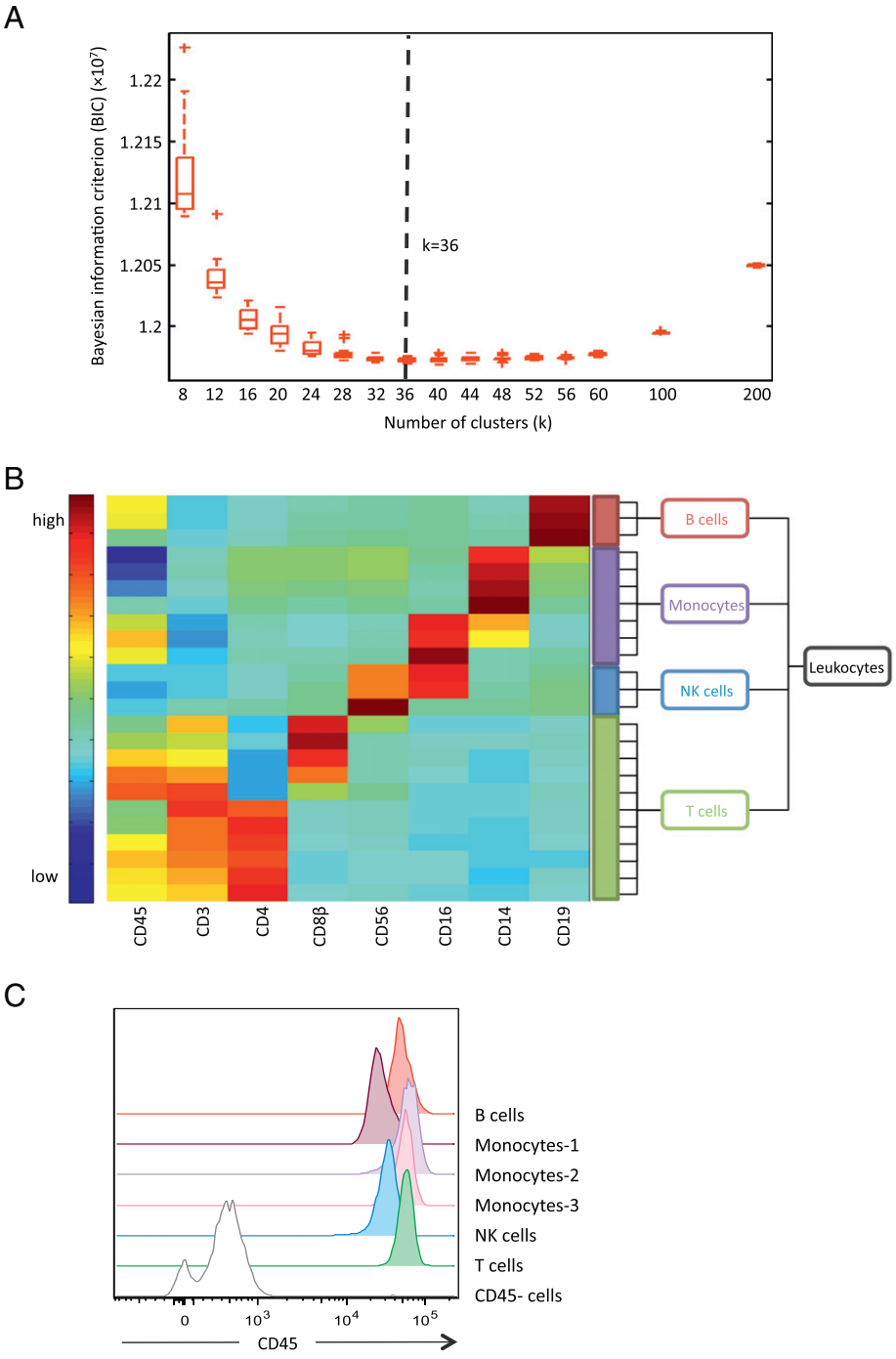
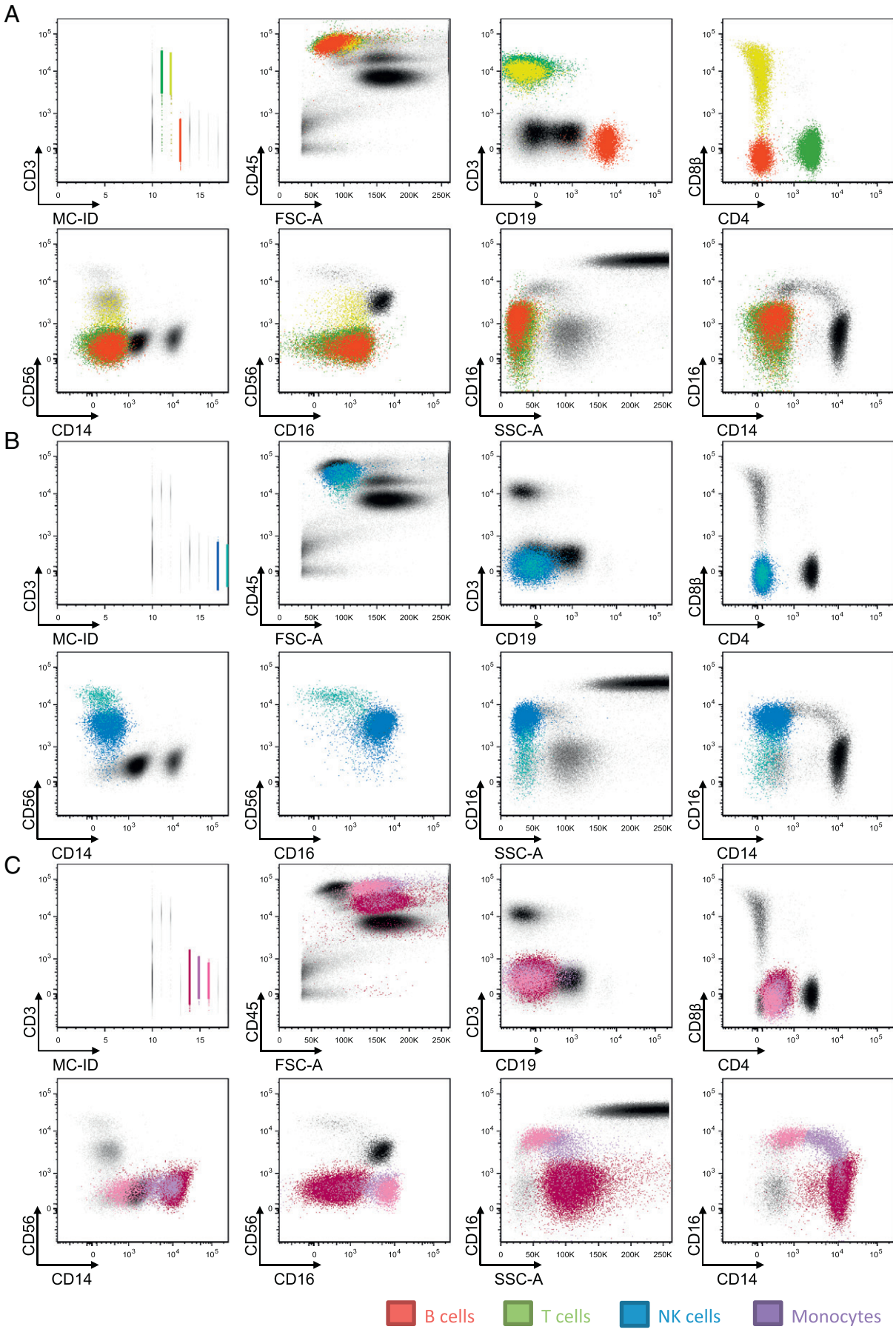


Figure 2 Number of clusters and mapping to cell types. (A) The number of clusters k is determined with the minimum average Bayesian Information Criterion (BIC) when evaluated on 20 random initial solutions for each choice of k. For the lineage panel, k = 36 is optimal. (B) User-based aggregation of FlowGM clusters into meta-clusters for immune cell type characterization with cluster centroid heat map (normalized coordinates). B cells are identified as CD19+, T cells are identified as CD3+ with two subsets: $CD4^+$ (T-1) and $CD8\beta^+$ (T-2), NK cells are identified as $CD56^+$ with two subsets: $CD16^{hi}$ (NK-1) and $CD56^{hi}$ (NK-2), monocytes are identified as three subsets: $CD14^{hi}$ (Mono-1), $CD14^{hi}CD16^{hi}$ (Mono-2) and $CD14^{lo}CD16^{hi}$ (Mono-3). The manually assigned cell types are indicated on the right. (C) Distribution of CD45 intensity for different cell types of interest in the reference donor.


A
B
C
CD3
MC-ID
CD45
FSC-A
CD3
CD19
CD8β
CD4
CD56
CD14
CD56
CD16
CD16
SSC-A
CD16
CD14
B cells
T cells
NK cells
Monocytes

3.2. Identification of the major cell lineages by FlowGM

We first applied FlowGM to the lineage panel dataset [Hasan et al., co-submission]. Cells were stained with the markers CD45, CD3, CD4, CD8β, CD14, CD16, CD19, and CD56. Following the approach of the manual analysis by Hasan et al., we used forward and side scatter (FSC/SSC) solely to exclude doublets; the remainder of our data analysis is performed on the dimensions of the indicated eight markers. The number of events in the data files ranged from 106,000 to 787,000. After filtering out doublets, FlowGM estimated the optimal number of clusters k to be 36, using the BIC (see Materials and methods, Section 2.4) on the reference donor (Fig. 2A).

Once k was determined, FlowGM performed EM clustering 100 times, starting with different random initial configurations of k clusters. The clustering solution with the highest likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$ constitutes the *reference clustering*, whose clusters were then manually labeled with the different cell types of interest (i.e., leukocyte subpopulations). The corresponding cluster centroids are represented as a heat map, with the assigned cell types indicated (Fig. 2B).

Note that only 24 of the 36 clusters corresponded to cell types of interest, and the color coding is chosen independently for each marker to resolve the entire spectrum of expression across these cell types (using the Matlab HeatMap function). For example, as $CD45^-$ cell populations were not of interest in this study, all selected cells were $CD45^+$ and as indicated by the normalization, the lowest and highest levels of CD45 expression were observed in monocytes and T cells, respectively (Figs. 2B, C).

To facilitate the understanding of our findings and permit user cross-validation, FlowGM allows the embedding of cluster IDs and meta-cluster IDs as additional channels (designated "C-ID" and "MC-ID", respectively) into the FCS input file, permitting importation of all data into FlowJo (or other FCS-compatible software). FlowJo visualizations of the labeled FlowGM lineage clusters confirmed our GMM-based assignments (Fig. 3). By gating on MC-ID to select one FlowGM meta-cluster, it is possible to view the clustered cells in 2D projections that correspond to manual gating strategies. FlowJo visualizations of all 36 FlowGM clusters are shown in Fig. S1. Backgating is also possible: starting with manual gated data and examining where the captured events cluster in C-ID or MC-ID space (data not depicted).

3.3. Pre-filtering supports clustering of rare dendritic cell subsets

We next evaluated the performance of the method on rare subsets of cells (<1% of the total cell events). In addition to the elimination of doublets early in the analysis, we identified the need for pre-filtering of cells considered by the user as uninteresting – similar to the use of a "Dump" gate – only in the case of FlowGM the procedure is automated and thus removes operator bias. Pre-filtering of the DC panel was based on a two-component, two-dimensional GMM that utilized data from CD14 and HLA-DR markers. Thresholds were automatically set at the 95th percentiles of the CD14/HLA-DR double-negative population (represented by the red line, Fig. 4A). The resultant cells were investigated using the FCS embedding feature of FlowGM, and inspection of representative files revealed accurate retention of desired $HLA\text{-}DR^+$ and/or $CD14^+$ cells (Fig. 4B).

Next, we estimated k using the BIC and defined a clustering solution using data from a reference donor (Fig. S2). Of the 40 clusters defined as the optimal fit, 22 were of interest and manual labeling of the meta-clustered data captured five myeloid cell subsets: cDC1, identified by their high BDCA2 MFI and low expression of CD14; pDCs, identified by the highest BDCA2 and BDCA4 MFIs; cDC3, identified by their expression of BDCA3; $CD14^{lo}$ monocytes, identified by the intermediate expression of CD14; and $CD14^{hi}$ monocytes, by the high CD14 MFI (Fig. S2B). Again, we highlight that the data represented in the heat map has been normalized, and in instances where all cell populations are positive for a given marker (i.e., HLA-DR), the normalization will scale values to span the range of marker expression. To illustrate the distributions of HLA-DR intensity, histogram plots for DCs and monocytes are shown (Fig. S2C).

Next, an initial post-filter removed dead cells from each meta-cluster, based on the Dump channel. A second post-filter removed cells from cDC1 and cDC3 populations based on expression of BDCA1 and BDCA3 respectively, of the CD14/HLA-DR double-negative population that was previously filtered out.

As a final validation step, we compared the level of marker expression between retained cells and events that were removed by the filtering process. Across all dimensions of the data set, we confirmed the efficacy of the pre-filtering approach (Figs. 4C, D). Additional visual confirmation can be found in the FlowJo-projected data, where meta-clustered data is overlaid on the total cell events in a representative file (Fig. S3).

3.4. FlowGM is robust to selection of reference donor and may be applied to uncompensated data

One potential concern with the FlowGM approach is the sensitivity of the clustering result to the choice of the reference sample in Step 1 (cf. Section 2.4). This is an important issue, as the resulting reference clustering will be used as the basis to cluster the data from all other samples. While practitioners may have a good intuition about which one of the input samples is "representative", the degree of sensitivity to this choice could, in principle, be large.

We therefore investigated whether a more representative reference clustering based on a larger group of samples would be needed. To this end, we constructed 11 different clusterings: the originally chosen reference clustering (which we denote here by 1*), and ten alternative reference clusterings

Figure 3 Visualization of labeled meta-clusters in FlowJo Cluster IDs is incorporated into the FlowJo input file. Shown are meta-clusters with all principal manual gating steps, starting with SSC-A/Meta-Cluster ID (MC-ID). (A) The identified $CD19^+$ B cells (red) and $CD4^+$ (green) and $CD8\beta^+$ (yellow) subsets of $CD3^+$ T cells. (B) $CD56^{hi}$ (light blue) and $CD16^{hi}$ (dark blue) NK cell sub-populations. (C) $CD14^{hi}$, monocytes (Mono-1, mauve) $CD14^{hi}CD16^{hi}$ monocytes (Mono-2, lavender) and $CD14^{lo}CD16^{hi}$ monocytes (Mono-3, light purple).

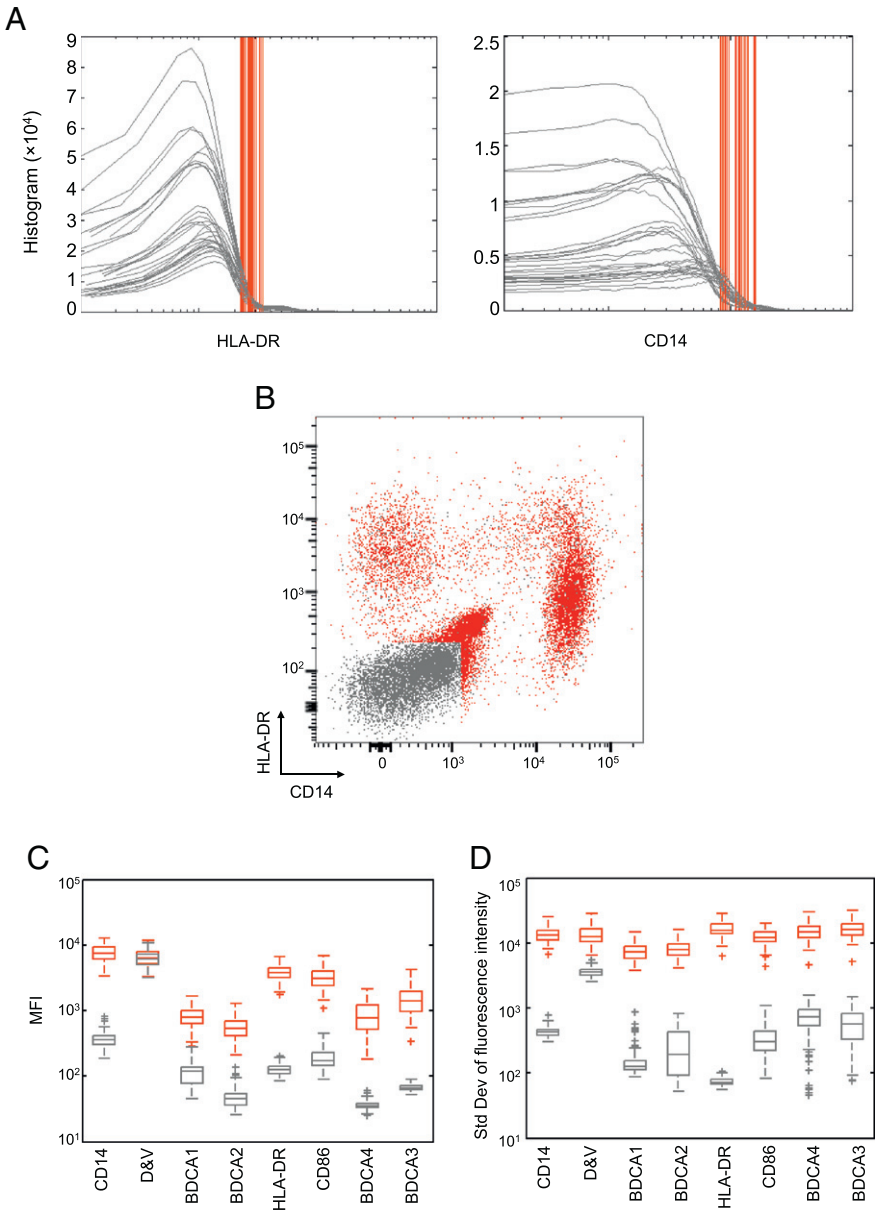


Figure 4 Pre-filtering for analysis of rare cell populations. (A) Pre-filtering in dendritic cells (DC) by low expression of CD14 and HLA-DR. Red lines indicate the thresholds that were automatically determined using GMM. (B) Validation of pre-filtering using FlowJo visualization. (C) MFI of filtered (gray) and remaining (red) cells. Pre-filtered cells display a lower MFI in all channels except Dump. (C) Standard deviation of fluorescence intensity for the same cell population. Filtered cells display less variation.

(1, … 10) of increasing complexity, which were obtained by selecting a series of 10 samples from randomly chosen donors, and then merging the samples 1, …, *i* for each *i* = 1, …, 10. Merging different samples without alignment can be expected to create reference clusterings that contain technical shifts, and thus could translate into significant variation in the clustering result.

For each possible pair of these 11 reference clusterings, we then determined the similarity of the two outcomes after clustering, using the F-measure [11,12] (Fig. 5A). Notably, the F-measure values were close to 1, independently, for all pairs of reference clusterings, indicating that the different reference clusterings did not translate into significantly different clustering outcomes. The locations of the resulting cell types for the different reference clusterings were further represented in parallel coordinate plots (Fig. 5B). Except for the Mono-2 and Mono-3 populations, all coordinates match extremely well among the different reference clusterings across all dimensions. Together, these observations suggest that the choice of the initial reference clustering may not have a large impact on the resulting outcome.

We also investigated the impact of compensation. Routinely, automatic hardware compensation [Hasan et al., co-submission] is employed. Here, we compare the results of our approach on the same input data in an uncompensated state; machine-compensated; or machine-compensated and FlowJo-corrected. The computational analyses on these three datasets were initialized with the re-estimated parameters from the

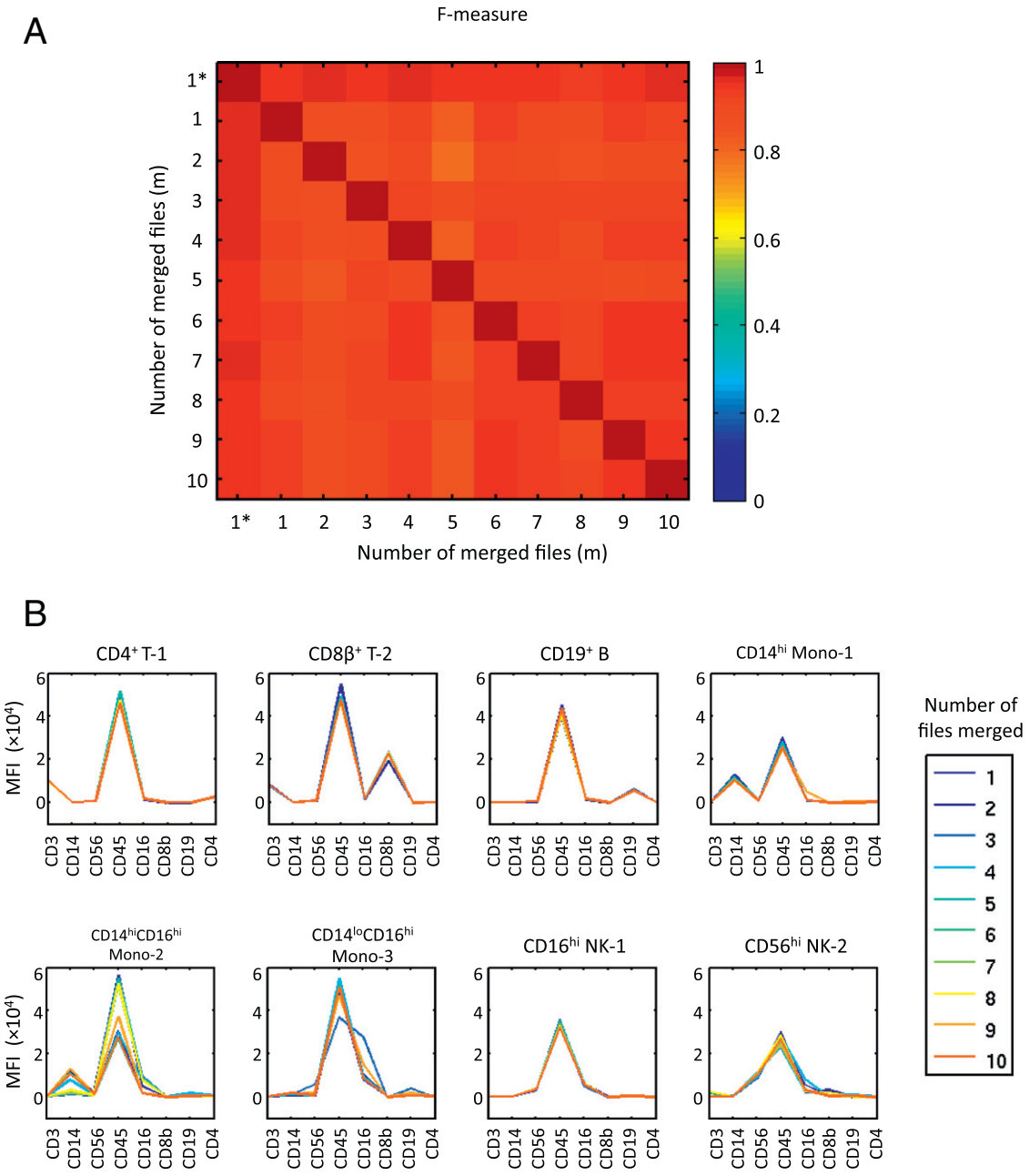


Figure 5 Differences in reference clustering do not impact cell type identification Different reference clusterings are generated by merging data from one to ten randomly selected donors; solutions are then applied to 115 cohort donors. (A) Pairwise average similarity (F-measure) of solutions over 115 cohort donors after using different reference clusterings. (B) Mean fluorescence intensity (MFI) of each identified cell population from different reference clusterings.

reference clustering on machine-compensated data. The counts for three repeatability samples obtained from different dataset are shown (Fig. S4), and indicate that FlowGM is insensitive to instrument compensation, and therefore resistant to potential compensation error in the context of large datasets.

3.5. Benchmarking of FlowGM demonstrates its reliability and utility

To directly compare FlowGM clusters to manually gated data sets, we first calculated, for each hand-gated cluster in the reference donor, the percentage of its events present in every other FlowGM cluster (Fig. 6A). The values indicated that, overall, the two approaches group events similarly. The one exception were monocytes, where FlowGM supported easy segregation of the $CD14^{hi}CD16^{hi}$ sub-population of monocytes (Mono-2) from $CD14^{lo}CD16^{hi}$ sub-monocytes (Mono-3), despite the lack of additional monocyte-specific markers (e.g., MCSF-1, CX_3CR1, CCR2 PMID: 20832340).

We also studied the variability of manual and FlowGM-derived cell counts across the repeatability samples studied in Hasan et al. (Fig. 6B). We find that FlowGM results showed

A

		FlowJo gates						
		CD4+	CD8+	CD19+	CD14^{hi} mono	CD16^{hi} mono	CD16^{hi} NK	CD56^{hi} NK
FlowGM meta-clusters	T - 1	99.8	0	0	0	0	0	0
	T - 2	0	98.7	0	0	0	0	0
	B	0	0	98.8	0	0	0	0
	Mono - 1	0	0	0	97.3	0	0	0
	Mono - 2	0	0	0	2.64	42.7	0	0
	Mono - 3	0	0	0	0	50.2	0	0
	NK - 1	0	0	0	0	3.8	96.2	1.5
	NK - 2	0	0	0	0	0	0.7	98.2
	Unmapped clusters	0.2	1.3	1.2	0.06	3.3	3.1	0.3

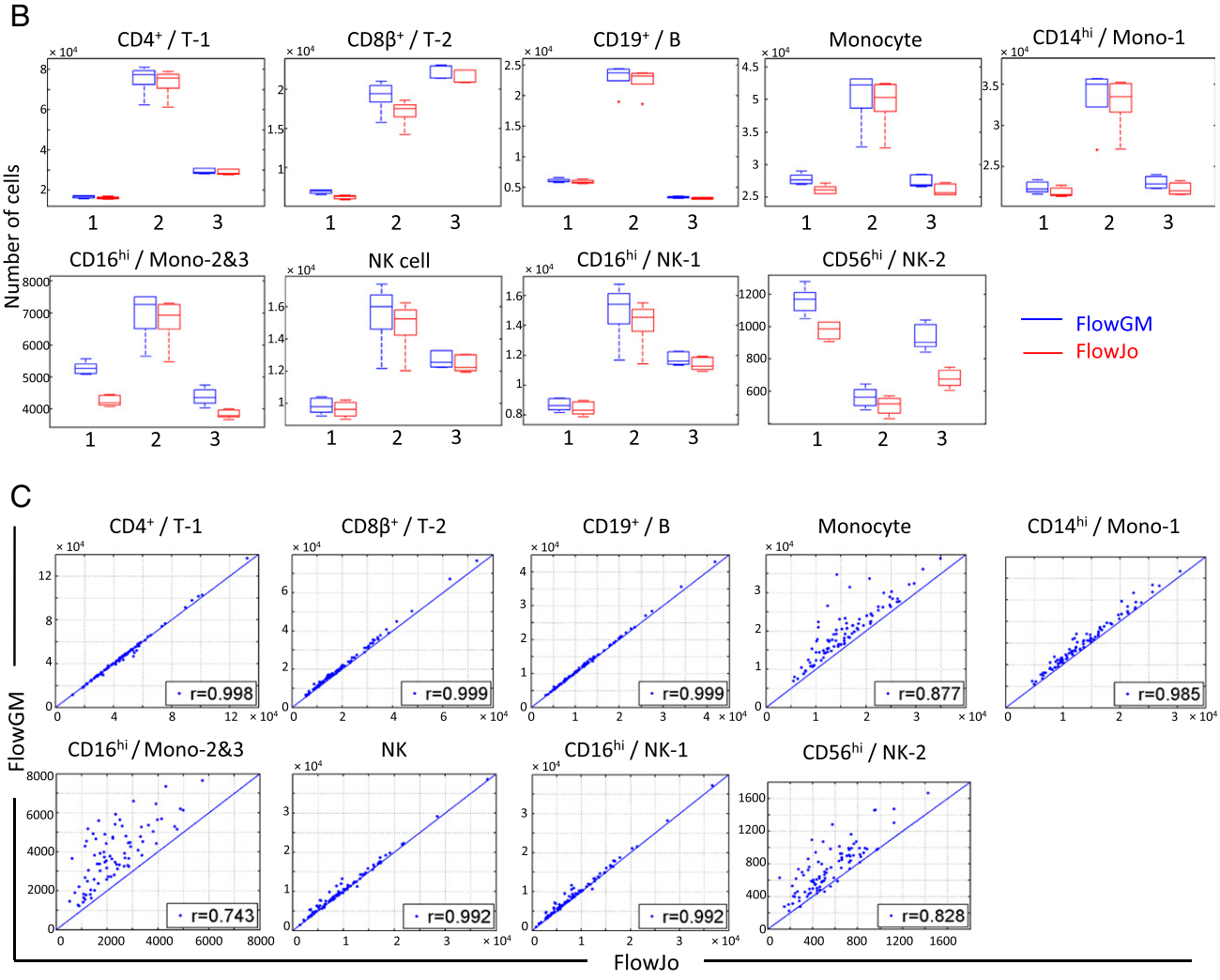


Figure 6 Comparison of manually gated data and FlowGM analysis. (A) Performance on reference donor: percentage of events in FlowJo cluster present in FlowGM clusters. (B) Performance on repeatability data: counts of each cell type for three donors with five replicates. The FlowGM results show a comparable CV with manually gated data. (C) Performance on 115 cohort donors: manually gated data and FlowGM analysis highly agree (r = 0.944) on 115 cohort donors.

good agreement with the results from manual analyses. The slight bias for higher numbers from FlowGM may stem from the need for high-dimensional information to confidently assign certain events to cell types (as in the schematic example shown, Fig. 1A). Coefficients of variation (CVs), which represent variation of data analysis and experimental variation, were at similar levels, further indicating the high accuracy of FlowGM analysis.

Table 1 Repeatability.

	Donor [a]:	#1	#2	#3
Lineage	CD4+ T cells	16870 (4.4) [b]	77306 (9.9)	28838 (4.4)
	CD8β + T cells	6986 (3.8)	19408 (10.6)	21416 (4.0)
	CD19+ B cells	5983 (5.3)	23679 (9.8)	3325 (4.1)
	Monocytes	27615 (3.0)	42233 (11.0)	26894 (3.5)
	$CD14^{hi}CD16^{lo}$ mono	22269 (3.2)	34969 (10.9)	22872 (3.2)
	$CD14^{hi}CD16^{hi}$ mono	3196 (4.2)	3759 (11.7)	1436 (8.9)
	$CD14^{lo}CD16^{hi}$ mono	2058 (3.3)	3505 (10.9)	2907 (5.3)
	NK cells	9803 (5.1)	15989 (12.9)	12534 (4.0)
	$CD16^{hi}$ NK	8633 (4.9)	15424 (13.0)	11632 (3.7)
	$CD56^{hi}$ NK	1171 (7.4)	565 (11.2)	902 (8.9)
T cell	CD4+ T cells	13172 (4.5)	64809 (16.4)	23450 (0.7)
	CD4+ T_N	3043 (4.8)	23398 (13.8)	8961 (8.1)
	CD4+ T_{CM}	8973 (4.4)	39350 (18.2)	13044 (3.6)
	CD4+ T_{EM}	1044 (6.7)	3329 (18.3)	1250 (11.4)
	CD8β + T cells	5245 (5.7)	14847 (16.8)	15283 (3)
	CD8β + T_N	553 (8.2)	5692 (16.8)	5903 (2.3)
	CD8β + T_{CM}	2297 (6.2)	5737 (13.6)	5996 (7.7)
	CD8β + T_{EM}	548 (10.2)	1181 (15)	1092 (21.2)
	CD8β + T_{EMRA}	717 (5.1)	1206 (46.8)	954 (16)
	CD8β + 27^{int}	1036 (8.7)	1096 (23.3)	1516 (11.7)
	CD4+ CD8α + T cells	153 (11.4)	770 (19.3)	539 (28)
DC	CD14+ monocytes	25232 (12.2)	29764 (4.4)	21287 (8.4)
	pDC	304 (18.5)	409 (4.1)	438 (5.0)
	cDC1	2159 (12.1)	5188 (3.9)	1677 (10.4)
	cDC3	42 (30)	87 (16)	44 (8.1)
PMN	Neutrophils	96062 (14.3)	188428 (13.0)	119529 (12.0)
	Basophils	1751 (11.4)	5878 (7.2)	2323 (11.6)
	Eosinophils	10483 (13.2)	18539 (10.6)	22329 (6.2)

[a] Fresh blood samples from three healthy donors were divided into five aliquots each and immediately stained using four antibody panels.
[b] Median absolute cell counts per 1 mL of blood in five independent analyses is represented for each cell population, as well as the corresponding coefficient of variation (CV).

Absolute counts and CVs for the repeatability data from all four panels are provided (Table 1). The estimation of the number of clusters and the resulting cluster positions, and assignments to cell types for the T cell and PMN panels are shown in Figs. S5 and S6 respectively. For the observed cell types, absolute counts were highly reproducible, with most CVs <15%. Compared to results of Hasan et al. [co-submission], the level of reproducibility of FlowGM was similar to the manual gating results across all four panels.

Finally, we used FlowGM-generated absolute cell counts of the lineage panel across 115 donors from the Milieu Intérieur cohort [Thomas et al., co-submission], comparing results to those obtained by manual gating. Again, results were highly concordant (Fig. 6C). The running time of the computational analysis for a single panel depends on the number n of measured events in each sample and the number k of clusters. For the panels analyzed here, the computation required 0.5 h (DC panel) and ~4 h (lineage panels) on a standard laptop PC.

4. Discussion

The FlowGM flow cytometry approach was developed to address the need for fast, robust and high-quality analysis for the Milieu Intérieur Consortium study. Our comprehensive validation study has shown that FlowGM has produced user-validated results whose quality is on par with, and in some cases, exceeds, the hand-gating approach. This is an exciting finding, as its simple computational approach does not require the expert knowledge and experience that is available to human operators. One important difference lies in the systematically higher number of events assigned to cell types by FlowGM, which suggests that the full dimensionality of the data, instead of two-dimensional views, allows for assigning cells that are unassigned in manual two-dimensional analysis due to the lacking dimensionality and user-bias. Another facet of this fundamental difference may be the observed ability of FlowGM to segregate subpopulations of monocytes without the need for an additional specific marker. Notably, separation of $CD14^{lo}CD16^{hi}$ monocytes from NK cells and other cell populations was achieved by integrating information from all eight dimensions.

When comparing the design of FlowGM workflow to other computational clustering approaches, a characteristic difference lies in the choice to computationally model single cell types as mixtures of Gaussians, as opposed to single Gaussians, or other distributions, coupled with the incorporation of knowledge and experience of a human operator to define which clusters belong to the same cell

type (referred to herein as meta-clusters). This design may constitute a 'sweet spot' in cytometry workflow design: A fast and efficient overall workflow, combined with a mathematical model that is flexible enough to model experimental data well, the solution of a hard core problem (the assignment of cell types to clusters) using operator intervention, and the limitation of this intervention to a single reference sample, as the transposition of this knowledge to all other samples can be automated with high accuracy.

The minimization of operator intervention means not only significant savings in terms of manual effort, but also the elimination of variability between different samples introduced by subjective decisions, and a considerable improvement in transparency and reproducibility of the path from the samples to the absolute and relative cell counts. Furthermore, the facility with which results are accessible for human inspection using conventional tools, and the relative simplicity of the FlowGM approach itself imply a high level of accessibility to non-specialists that – we believe – will continue to play an important role in the evolution of the approach.

We believe that the FlowGM workflow is applicable to most other flow cytometry datasets, and anticipate that the need for fast, robust, and high-quality analysis of large cytometry datasets will only increase. Adaptations of the method may be required for heterogeneous samples, in which no single reference sample may be representative for all others, or in cases where certain subpopulations may be activated (e.g., disease populations). We believe that there are relatively straightforward approaches to extend FlowGM to automatically detect cases of inadequate fit, for example, through the introduction of additional reference donors (with recursive iteration of the manual Step 5). The increased availability of experimental datasets that have been acquired under standardized conditions may facilitate comparison and integration, which may lead to the necessary insights and technical developments to fully automate flow cytometry data analysis.

5. Conflict of interest statement

The authors declare that there are no conflicts of interest.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.clim.2014.12.009.

Acknowledgments

Consortia
The Milieu Intérieur Consortium[1] is composed of the following team leaders: Laurent Abel (Hôpital Necker), Andres Alcover, Philippe Bousso, Pierre Bruhns, Ana Cumano, Marc Daëron, Cécile Delval, Caroline Demangel, Ludovic Deriano, James Di Santo, Françoise Dromer, Gérard Eberl, Jost Enninga, Odile Gelpi, Antonio Freitas, Ivo Gomperts-Boneca, Serge Hercberg (Université Paris 13), Olivier Lantz (Institut Curie), Claude Leclerc, Hugo Mouquet, Sandra Pellegrini, Stanislas Pol (Hôpital Côchin), Lars Rogge, Anavaj Sakuntabhai, Olivier Schwartz, Benno Schwikowski, Spencer Shorte, Vassili Soumelis (Institut Curie), Frédéric Tangy, Eric Tartour (Hôpital Européen George Pompidou), Antoine Toubert (Hôpital Saint-Louis), Marie-Noëlle Ungeheuer, Lluis Quintana-Murci[2], Matthew L. Albert[3].

Additional information can be found at: http://www.pasteur.fr/labex/milieu-interieur.

References

[1] V. Orru, M. Steri, G. Sole, C. Sidore, F. Virdis, M. Dei, S. Lai, M. Zoledziewska, F. Busonero, A. Mulas, M. Floris, W.I. Mentzen, S.A. Urru, S. Olla, M. Marongiu, M.G. Piras, M. Lobina, A. Maschio, M. Pitzalis, M.F. Urru, M. Marcelli, R. Cusano, F. Deidda, V. Serra, M. Oppo, R. Pilu, F. Reinier, R. Berutti, L. Pireddu, I. Zara, E. Porcu, A. Kwong, C. Brennan, B. Tarrier, R. Lyons, H.M. Kang, S. Uzzau, R. Atzeni, M. Valentini, D. Firinu, L. Leoni, G. Rotta, S. Naitza, A. Angius, M. Congia, M.B. Whalen, C.M. Jones, D. Schlessinger, G.R. Abecasis, E. Fiorillo, S. Sanna, F. Cucca, Genetic variants regulating immune cell levels in health and disease, Cell 155 (2013) 242–256.
[2] FlowJo, TreeStar Software, Ashland, OR.
[3] X. Hu, H. Kim, P.J. Brennan, B. Han, C.M. Baecher-Allan, P.L. De Jager, M.B. Brenner, S. Raychaudhuri, Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer T cells, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 19030–19035.
[4] AdiCyt, Adinis S.R.O., Bratislava, Slovenia, 2013.
[5] N. Aghaeepour, R. Nikolic, H.H. Hoos, R.R. Brinkman, Rapid cell population identification in flow cytometry data, Cytometry A 79 (2011) 6–13.
[6] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.
[7] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, J. R. Stat. Soc. Ser. B 39 (1977) 1–38.
[8] G.E. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[9] Matlab and Statistics Toolbox Release 2012b, The MathWorks Inc., Natick, Massachusetts, United States, 2012.
[10] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.
[11] B. Ellis, P. Haaland, F. Hahne, N.L. Meur, N. Gopalakrishnan, J. Spidlen, flowCore: Basic structures for flow cytometry data, R package version 1.30.7 (2014).
[12] C.J. Van Rijsbergen, Information Retrieval, 2nd ed., 1979. (Butterworth).
[13] N. Aghaeepour, G. Finak, C.A.P.C. Flow, D. Consortium, H. Hoos, T.R. Mosmann, R. Brinkman, R. Gottardo, R.H. Scheuermann, Critical assessment of automated flow cytometry data analysis techniques, Nat. Methods 10 (2013) 228–238.

[1] Unless otherwise indicated, partners are located at Institut Pasteur, Paris.

[2] Co-coordinator of the Milieu Intérieur Consortium.

[3] Co-coordinator of the Milieu Intérieur Consortium.